DetMatch: Two Teachers are Better Than One for Joint 2D and 3D Semi-Supervised Object Detection

Jinhyung Park[1*], Chenfeng Xu[2†], Yiyang Zhou[2], Masayoshi Tomizuka[2], and Wei Zhan[2]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
jinhyun1@andrew.cmu.edu
[2] University of California, Berkeley, Berkeley CA 94720, USA
{xuchenfeng, yiyang.zhou, tomizuka, wzhan}@berkeley.edu

Abstract. While numerous 3D detection works leverage the complementary relationship between RGB images and point clouds, developments in the broader framework of semi-supervised object recognition remain uninfluenced by multi-modal fusion. Current methods develop independent pipelines for 2D and 3D semi-supervised learning despite the availability of paired image and point cloud frames. Observing that the distinct characteristics of each sensor cause them to be biased towards detecting different objects, we propose DetMatch, a flexible framework for joint semi-supervised learning on 2D and 3D modalities. By identifying objects detected in both sensors, our pipeline generates a cleaner, more robust set of pseudo-labels that both demonstrates stronger performance and stymies single-modality error propagation. Further, we leverage the richer semantics of RGB images to rectify incorrect 3D class predictions and improve localization of 3D boxes. Evaluating our method on the challenging KITTI and Waymo datasets, we improve upon strong semi-supervised learning methods and observe higher quality pseudo-labels. Code will be released here: https://github.com/Divadi/DetMatch.

Keywords: Semi-Supervised Learning, Multi-Modal Learning, Object Detection.

1 Introduction

Recent advances in Semi-Supervised Learning (SSL) for object recognition focus on the single-modality setting, demonstrating improvements in either 2D or 3D detection when leveraging unlabeled samples of that modality. However, SSL works rarely study the combination of 2D and 3D sensors. In recently published

[*] Work conducted during visit to University of California, Berkeley.
[†] Corresponding author

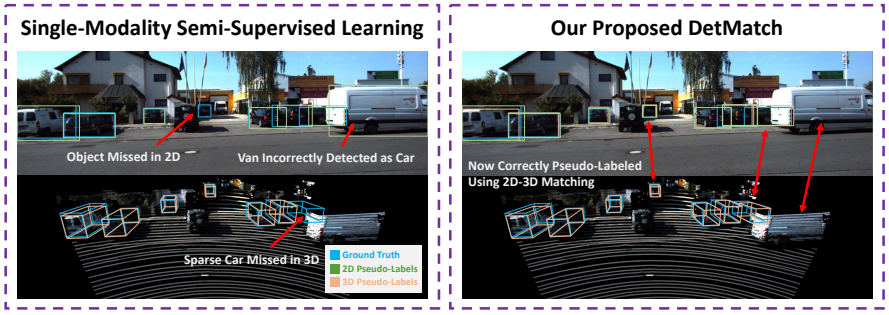

Fig. 1: Matching 2D and 3D detections, DetMatch removes false negatives and positives to generate cleaner pseudo-labels. Points are colored for visualization

datasets, autonomous vehicles are equipped with a comprehensive collection of sensors that yields multi-modal observations of each scene. Among these devices, 2D RGB cameras and 3D LiDARs have emerged as two independently useful but also mutually complementary modalities. Thus, it is important for SSL methods to utilize both 2D and 3D modalities for autonomous driving applications.

We propose a novel multi-modal SSL framework, **DetMatch**, that leverages paired but unlabeled data of multiple modalities to train stronger single-modality object detectors. Our pipeline is agnostic to the designs of the detectors, allowing for flexible usage in conjunction with perpendicular advancements in architectures. Further, by yielding single-modality models, DetMatch does not constrain the trained detectors to the multi-modal or even the autonomous driving setting.

We observe that differences in modality characteristics between RGB images and point clouds cause them to each be better at detecting different types of objects as illustrated in Figure 1. 3D point clouds are sparse, and their lack of color causes structurally similar objects to be indistinguishable. On the other hand, 2D RGB images contain a dense array of color information, allowing for easier discrimination of similarly shaped classes and better detection of objects with few 3D points captured. However, unlike point clouds, RGB images lack depth values. Each point in the 3D point cloud represents an exact, observed location in 3D space, making objects spatially separable - this facilitates 3D detection of objects that have overlapping, similar-colored projections in 2D. These factors support our intuition that not only are RGB images and point clouds mutually beneficial, but that their detection results are strongly complementary.

To leverage this relationship for SSL while keeping each detection model single-modal, we associate 2D and 3D results at the detection level. Since 2D and 3D have their own strengths, we use predictions in each modality that have a corresponding detection in the other modality to generate a cleaner subset of box predictions that is used to pseudo-label the unlabeled data for that modality. We find that such pseudo-labels chosen using multiple modalities outperform single-modality generated pseudo-labels. Although this method exploits the advantages of each modality to generate stronger pseudo-labels, it insufficiently utilizes the RGB images' unique rich semantics. In the previous pipeline, a correctly localized

& classified 2D detection cannot directly rectify a poor 3D detection. To remedy this gap, we additionally enforce box and class consistency between matched 2D pseudo-labels and 3D predictions and observe improved performance.

Our main contributions are as follows:

- We observe that differences in characteristics between 2D and 3D modalities allow objects of high occlusion to be better detected in 3D, and objects of similar shape but different class to be better identified and localized in 2D.
- Our SSL framework leverages the mutually beneficial relationship between multiple modalities during training to yield stronger single-modality models.
- We extensively validate DetMatch the difficult KITTI [12] and Waymo [57] datasets, notably achieving around 10 mAP absolute improvement over labeled-only 3D baseline on the 1% and 2% KITTI settings and a 10.6 AP improvement for Pedestrians in 3D on the 1% Waymo setting.

2 Related Work

Semi-Supervised Learning. SSL methods either use consistency regularization [1, 24, 43, 47, 60] or pseudo-labeling [2, 3, 25, 53, 77]. The former forces noised predictions on unlabeled images to be consistent. The seminal work [1] enforces consistency over dropout, Temporal Ensembling [24] stores exponential moving averages (EMA) of past predictions, and Mean Teacher [60] enforces consistency between "student" and "teacher" models, the latter an EMA of the former.

Pseudo-labeling methods explicitly generate labels on unlabeled data and train on them in lieu of ground truth. MixMatch [3] ensembles over augmentations, ReMixMatch [2] uses weak augmentations for labeling and strong augmentations for training, and FixMatch [53] uses a confidence threshold to generate labels. Our method builds on intuitions from Mean Teacher [60] and asymmetric augmentations [3, 27, 33, 53] to ensure the teacher model can correctly supervise the student by maintaining an advantage over the student.

SSL for Object Detection. 2D detection models [5, 29, 32, 44, 45, 61] consist of a feature extraction backbone [14], a region proposal network [32, 45], and optionally, a second-stage proposal refinement module [5,45]. 3D object detection methods [13, 28, 48, 72, 82] follow a similar structure, instead using voxel [9, 11, 13, 50, 72] or point [39, 41, 49, 70, 74] representations instead of 2D modules. Our proposed DetMatch is agnostic to the single-modality detectors used.

Some 2D SSL object detection methods [19,59] enforce consistency over augmentations, STAC [54] generates pseudo-labels offline, and Instant-Teaching [81] uses Mosaic [4] and MixUp [78]. A line of work [26, 71] improves thresholding, and others use EMA for predictions [73] and teacher models [33, 59]. Similarly, for 3D SSL, SESS [80] trains consistency over asymmetric augmentations and 3DIoUMatch [64] thresholds on predicted IoU. Compared to 2D, more 3D methods use offline labeling [6,42,65], with some [42,65] using ensembling and multiple timesteps to refine detections. Improvements in multi-frame fusion are perpendicular to our work, as our DetMatch generates cleaner per-frame pseudo-labels

that can be used for downstream multi-timestep aggregation and refinement. Unlike these single-modality SSL methods, our pipeline jointly leverages the unique characteristics of RGB and point clouds to improve SSL for each modality.

2D-3D Multi-Modal Learning. Many works have explored 2D-3D fusion for detection and segmentation. Some methods [23, 40, 66] constrain the 3D search space through 2D detection, while others fuse 2D and 3D features [16, 52, 63, 76, 79] or predictions [37, 38, 62, 68, 75]. Some works have explored cross modal distillation [8], contrastive pretraining [34–36], or directly transferring 2D model into 3D [69]. Most relevant to our work is xMUDA [18], which proposes a cross-modality loss for semantic segmentation domain adaptation. Their 3D model is supervised by 2D segmentation results and vice versa. However, unlike pixels and points on which segmentation is done, detections in 2D and 3D do not have a directly calculable bijective mapping, making cross-modal supervision in object detection a less constrained problem. Further, training box regression requires extra consideration. We address these difficulties in our framework.

3 Method

3.1 Problem Definition

In semi-supervised object detection, we have a small set of labeled data $\{(\mathbf{x}_i^l, \mathbf{y}_i^l)\}_{i=1}^{N_l}$ and a larger set of unlabeled data $\{\mathbf{x}_i^u\}_{i=1}^{N_u}$, where N_l and N_u are the number of labeled and unlabeled frames, respectively. We typically have $N_u >> N_l$. We omit the scripts on $\mathbf{x}_i^l$ when they are clear from context. In autonomous driving [12, 57] and indoor scene understanding [10, 17, 51, 56, 67], a single input sample is a multi-modal tuple $\mathbf{x} = (\mathbf{x}_{2D}, \mathbf{x}_{3D})$. $\mathbf{x}_{2D}$ is a 2D RGB image and $\mathbf{x}_{3D}$ is a 3D point cloud. Similarly, each ground truth annotation is a tuple of 2D and 3D labels, which in turn are each a set of boxes and classification labels:

$$\mathbf{y} = \left(\mathbf{y}_{2D} = \left\{(\mathbf{b}_{2D}, \mathbf{c}_{2D})^{(j)}\right\}, \mathbf{y}_{3D} = \left\{(\mathbf{b}_{3D}, \mathbf{c}_{3D})^{(j)}\right\}\right)$$

$\mathbf{b}_{2D} \in \mathbb{R}^4$ is a 2D box, $\mathbf{b}_{3D} \in \mathbb{R}^7$ is a 3D box, and $\mathbf{c} \in \{0, 1\}^C$ is a one-hot label indicating one of C classes. To reduce the labeling burden for training, we generate $\mathbf{y}_{2D}$ from $\mathbf{y}_{3D}$ by projecting $\mathbf{b}_{3D}$ to 2D to get $\mathbf{b}_{2D}$ using camera parameters. Thus, our pipeline requires no 2D labels for the target dataset.

3.2 Teacher-Student Framework

We use a student model $\mathbf{S}$ and a teacher model $\mathbf{T}$ of the same architecture. At a high level, the teacher $\mathbf{T}$ generates pseudo-labels on the unlabeled data that the student $\mathbf{S}$ trains on. For the teacher to correctly and stably supervise the student, the teacher must maintain an advantage over the student in terms of the performance. We accomplish this by iteratively updating and improving the teacher model through training via exponential moving average (EMA) accumulation:

$$\theta_{\mathbf{T}} \leftarrow \alpha\theta_{\mathbf{T}} + (1 - \alpha)\theta_{\mathbf{S}} \tag{1}$$

where α is the EMA momentum, and the θ are the model parameters. Unlike methods that pseudo-label offline [6, 42, 54], our student and its EMA teacher allow for continuous improvement of pseudo-labels throughout training.

3.3 Single-Modality Semi-Supervised Learning

Overview. In this section, we outline a straightforward teacher-student, single-modality SSL approach based on the state-of-the-art 2D SSL method Unbiased Teacher [33]. We find that with a well-tuned confidence threshold, this simple baseline compares favorably against more complicated approaches in 3D such as 3DIoUMatch [64]. We omit modality indicators $2D$ and $3D$ for this section, because this SSL baseline is applicable to any detection model.

Pre-training. For the teacher to reasonably guide the student from the start, we first pre-train the student model on the labeled data. Let $\mathbf{T}(\mathbf{x}) = \hat{\mathbf{y}}_\mathbf{T} = \left\{(\hat{\mathbf{b}}_\mathbf{T}, \hat{\mathbf{c}}_\mathbf{T})^{(j)}\right\}$ and $\mathbf{S}(\mathbf{x}) = \hat{\mathbf{y}}_\mathbf{S} = \left\{(\hat{\mathbf{b}}_\mathbf{S}, \hat{\mathbf{c}}_\mathbf{S})^{(j)}\right\}$ denote the predictions of the teacher and the student models respectively, with each consisting of a set of bounding boxes and classification probabilities. The loss on labeled samples is:

$$\mathcal{L}^l = \mathcal{L}_{loc}\left(\hat{\mathbf{y}}^l_\mathbf{S}, \left\{\mathbf{b}^{l(j)}\right\}\right) + \mathcal{L}_{cls}\left(\hat{\mathbf{y}}^l_\mathbf{S}, \left\{\mathbf{c}^{l(j)}\right\}\right) \tag{2}$$

where $\mathcal{L}_{loc}$ and $\mathcal{L}_{cls}$ represent the localization and classification losses, respectively. After the student is pre-trained to convergence, the teacher is initialized with the student weights before the SSL training begins.

Semi-Supervised Training. To retain representations learned from the labeled data, we train using an equal number of labeled and unlabeled samples per batch:

$$\mathcal{L} = \mathcal{L}^l + \lambda\mathcal{L}^u \tag{3}$$

where $\mathcal{L}^l$ is as defined in Equation 2, $\mathcal{L}^u$ is the loss on unlabeled samples and λ is a weighting hyperparameter. To train on unlabeled data, we get box predictions from the teacher and only keep the ones with maximum classification confidence above a threshold τ as pseudo-labels. We can write the teacher's generated pseudo-labels on the unlabeled data as:

$$\hat{\mathbf{y}}^{(>\tau)}_\mathbf{T} = \left\{(\hat{\mathbf{b}}_\mathbf{T}, \hat{\mathbf{c}}_\mathbf{T})^{(j)}\right\}^{(>\tau)} = \left\{\left(\hat{\mathbf{b}}^{(j)}_\mathbf{T}, \hat{\mathbf{c}}^{(j)}_\mathbf{T}\right) \;\middle|\; \max(\hat{\mathbf{c}}^{(j)}_\mathbf{T}) > \tau\right\} \tag{4}$$

giving us the unlabeled loss:

$$\mathcal{L}^u = \mathcal{L}_{loc}\left(\hat{\mathbf{y}}^u_\mathbf{S}, \left\{\mathbf{b}^{(j)}_\mathbf{T}\right\}^{(>\tau)}\right) + \mathcal{L}_{cls}\left(\hat{\mathbf{y}}^u_\mathbf{S}, \left\{\text{argmax}(\mathbf{c}^{(j)}_\mathbf{T})\right\}^{(>\tau)}\right) \tag{5}$$

After SSL training, we take the teacher as our final model for more stability.

Asymmetric Data Augmentation. Although EMA makes the teacher more stable than the student, EMA alone does not give the teacher a large enough advantage in performance. To further decouple their predictions, we adopt asymmetric data augmentation on the inputs of the teacher and the student. We use weak augmentation $\mathcal{A}_{weak}(\mathbf{x})$ for the teacher and strong augmentation $\mathcal{A}_{strong}(\mathbf{x})$ for the student. We find that this single-modality SSL framework outperforms 3DIoUMatch on driving datasets, so we adopt it as our baseline for comparison.

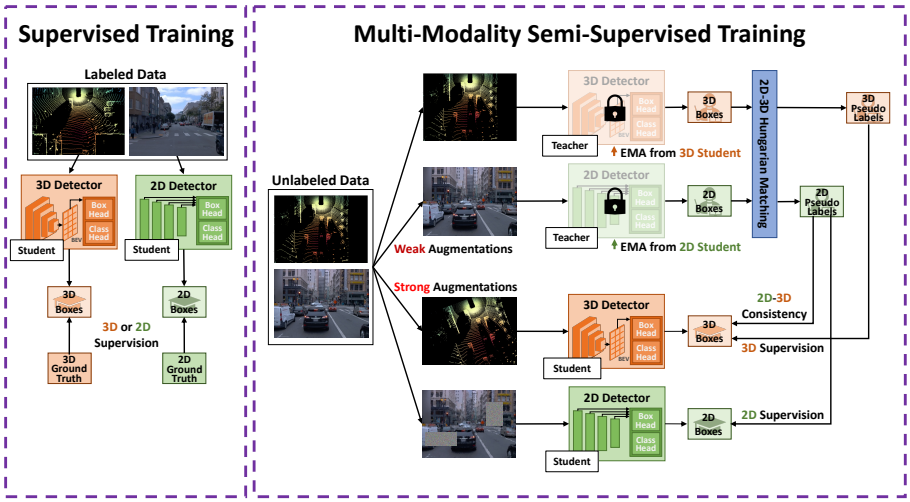

Fig. 2: The proposed DetMatch. We have a teacher and student for each modality and match 2D and 3D teacher predictions to supervise the students. The 2D teacher also directly supervises the 3D student through 2D-3D Consistency

3.4 Multi-Modality Semi-Supervised Learning

Overview. Although this single-modality SSL framework improves over labeled-only training, it has several disadvantages. Firstly, it does not leverage the paired 2D and 3D inputs, leading to sub-optimal single-modality results. Secondly, classification confidence is a poor measure of box localization performance as noted by prior work [20,55]. Finally, we find that single-modality self-training is prone to error propagation, leading to decreased performance in some cases.

To address these problems, we present our multi-modal semi-supervised framework shown in Figure 2. DetMatch jointly maintains a teacher and a student for each modality and matches 2D and 3D teacher predictions to generate a cleaner set of pseudo-labels. Furthermore, to leverage the unique advantages of dense, colorful 2D RGB images, we propose a 2D-3D consistency module that forces 3D student predictions to be similar to 2D teacher boxes. Our multi-modal framework also performs pre-training and keeps labeled losses $\mathcal{L}^l_{2D}, \mathcal{L}^l_{3D}$ during SSL training for each modality as in Section 3.3. As the pseudo-label generation changes, our unlabeled losses $\mathcal{L}^u_{2D}, \mathcal{L}^u_{3D}$ are different from Equation 5. We also introduce an additional $\mathcal{L}_{consistency}$ loss. The overall loss for our DetMatch is:

$$\mathcal{L} = (\mathcal{L}^l_{2D} + \mathcal{L}^l_{3D}) + (\mathcal{L}^u_{2D} + \mathcal{L}^u_{3D}) + \mathcal{L}_{consistency} \tag{6}$$

2D-3D Hungarian Matching & Supervision. A drawback of the pipeline in Section 3.3 is its use of classification confidence to determine pseudo-labels. We visualize this issue in the left plot of Figure 3, which shows that many 3D boxes with a low max score are highly overlapped with a ground truth box. Moreover, although scoring modules directly supervised by true IoU [48,64] are better than

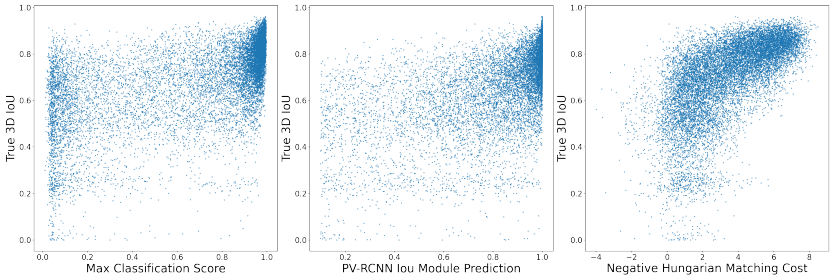

Fig. 3: Comparison between boxes' true 3D ground-truth IoU and various methods of assessing box quality on KITTI 1% unlabeled data

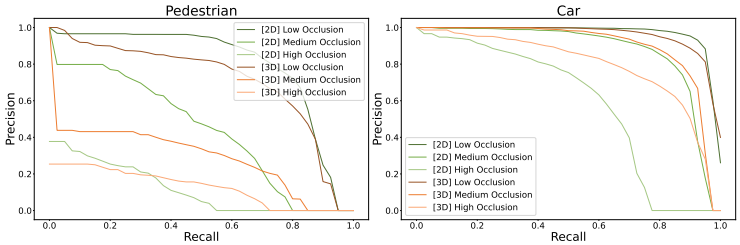

Fig. 4: 2D and 3D model performance at various occlusion levels

max classification score as shown in the middle plot, this IoU prediction module is unable to differentiate among high IoU values 0.6 - 0.9 as evidenced by the vertical cluster on the right side. As such, pseudo-labels generated using these single-modality measures of box quality prediction remain noisy.

We first examine the pros and cons of the 2D and 3D modalities. We plot in Figure 4 the P/R curves of 2D and 3D detections for Pedestrian and Car classes on the KITTI validation dataset, with a separate curve for ground truth objects labeled as low, medium, and high occlusion. We find that at the same occlusion level, 2D better detects and localizes Pedestrians when compared to 3D. Due to the sparsity of point clouds and their lack of color information, Pedestrians are often confused with poles and trees of similar shape in 3D. However, such ambiguous objects are clearly identifiable in the dense 2D RGB image.

On the other hand, 2D detection struggles with highly overlapping objects due to its lack of depth information - when viewed in the 3D point cloud, such overlapping objects are clearly separated. This trend is especially clear when viewing the P/R curves for the Car class. Although 2D outperforms 3D for objects of low occlusion, we see a clear reversal for highly occluded objects. These observations clearly demonstrate that 2D and 3D modalities are complementary at the detection level - a relationship we propose to leverage for SSL by choosing as pseudo-labels detections with a corresponding match in the other modality.

More specifically, as shown in Figure 5, we compute an optimal bipartite matching between 2D and 3D teacher predictions using the Hungarian Algorithm [22] and consider pairs with a matching cost below a threshold τ_{hung} "matched".

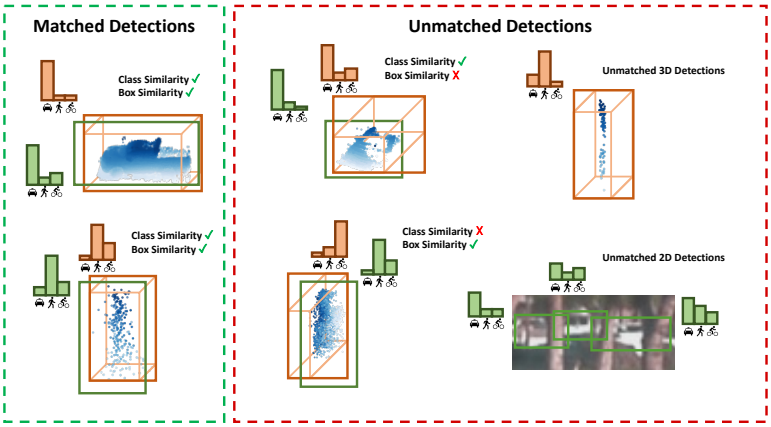

Fig. 5: Illustration of the 2D-3D Hungarian matching algorithm

The algorithm for matched pairs generation can be written as:

$$\left\{\left((\hat{\mathbf{b}}_\mathbf{T}, \hat{\mathbf{c}}_\mathbf{T})^{2D}, (\hat{\mathbf{b}}_\mathbf{T}, \hat{\mathbf{c}}_\mathbf{T})^{3D}\right)^{(j)}\right\}^{(<\tau_{hung})} = Hungarian_{2D\text{-}3D}^{\tau_{hung}}(\hat{\mathbf{y}}_\mathbf{T}^{2D}, \hat{\mathbf{y}}_\mathbf{T}^{3D}) \quad (7)$$

We omit notation for the matching algorithm and thresholding for brevity. Inspired by recent works [7,58] on detection using learnable queries, our matching cost between a pair of 2D and 3D box predictions has three components:

$$\mathcal{L}_{match}\left((\hat{\mathbf{b}}_\mathbf{T}, \hat{\mathbf{c}}_\mathbf{T})^{2D}, (\hat{\mathbf{b}}_\mathbf{T}, \hat{\mathbf{c}}_\mathbf{T})^{3D}\right) = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{iou}\mathcal{L}_{iou} + \lambda_{d\text{-}focal}\mathcal{L}_{d\text{-}focal} \quad (8)$$

Note that unlike classification score, a *lower* cost indicates a stronger match.

$\mathcal{L}_{L1}$ and $\mathcal{L}_{iou}$ are box consistency costs between the projected 3D box and the 2D box. To get the former, we project the 8 corners of the 3D box to the image and compute a tightly fitted 2D box. $\mathcal{L}_{L1}$ calculates l_1 loss between the 2D box parameters and $\mathcal{L}_{iou}$ calculates generalized IoU loss [46]. These costs force paired 2D and 3D pseudo-labels to refer to the same object agree on its localization. Unlike single-modality box localization confidence methods that suffer from modality-specific drawbacks and self-confidence bias, our multi-modal box consistency cost gives us a natural way to assess box quality.

$\mathcal{L}_{d\text{-}focal}$ calculates class prediction consistency between the 2D and 3D predictions. We formulate a double-sided version of FocalLoss [30]:

$$\mathcal{L}_{d\text{-}focal} = FocalLoss\left(\hat{\mathbf{c}}_\mathbf{T}^{2D}, \text{argmax}(\hat{\mathbf{c}}_\mathbf{T}^{3D})\right) + FocalLoss\left(\hat{\mathbf{c}}_\mathbf{T}^{3D}, \text{argmax}(\hat{\mathbf{c}}_\mathbf{T}^{2D})\right) \quad (9)$$

Note that this double-sided FocalLoss allows for a smooth trade-off between 2D and 3D confidence. A low-confidence 3D box *can still be chosen as a pseudo-label* if its matched 2D box has high confidence. Intuitively, high-confidence predictions of one modality can "promote" low-confidence predictions of the other modality, a dynamic selection not possible with simple confidence thresholding.

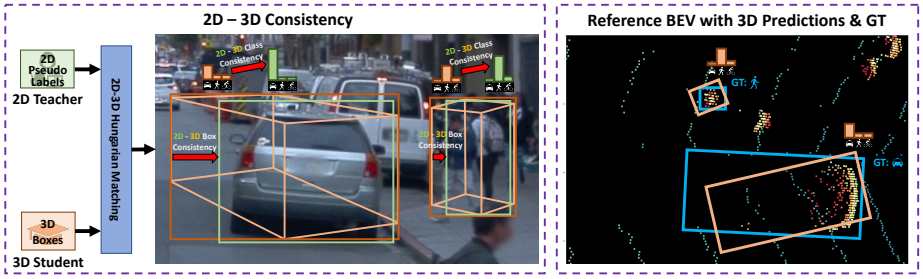

Fig. 6: The box and class consistency between the 2D teacher and the 3D student

Further, although this formulation of $\mathcal{L}_{focal}$ does prefer higher-confidence boxes, its motivation is different from that of confidence thresholding - $\mathcal{L}_{focal}$ considers *consistency* between classification predictions in 2D and 3D. If both modalities agree on the semantic class of a region, they will have a lower matching cost.

Our proposed 2D-3D matching cost is a remarkably more accurate measure of box localization quality as shown in the rightmost plot of Figure 3. We then use the matched and thresholded pairs of 2D and 3D teacher boxes as pseudo-labels to supervise the 2D and 3D students on the unlabeled data:

$$\begin{aligned}\mathcal{L}^{u}_{modal} =& \mathcal{L}_{loc}\left(\hat{\mathbf{y}}^{modal}_{\mathbf{S}}, \left\{(\mathbf{b}^{modal}_{\mathbf{T}})^{(j)}\right\}^{(<\tau_{hung})}\right) \\ &+ \mathcal{L}_{cls}\left(\hat{\mathbf{y}}^{modal}_{\mathbf{S}}, \left\{\text{argmax}\left((\mathbf{c}^{modal}_{\mathbf{T}})^{(j)}\right)\right\}^{(<\tau_{hung})}\right) \\ &\text{for } modal \in \{2D, 3D\} \end{aligned} \tag{10}$$

2D-3D Consistency. Through our 2D-3D Hungarian Matching, we generated a cleaner set of pseudo-labels to supervise each student. However, although we have leveraged the advantages 3D can provide 2D, we have not fully exploited the benefits 3D can get from 2D. We have fulfilled the former because although a core advantage of 3D is detection of highly occluded or visually unclear boxes, we need to differentiate these beneficial 3D teacher boxes from the false positives 3D detection is especially prone to. So, it is necessary to first match 3D boxes with 2D teacher boxes to filter noisy boxes while retaining the beneficial boxes.

On the other hand, the semantically rich format of 2D RGB images make class confusion less likely and instead enables better localization of non-heavily occluded objects as shown in Figure 4. So, high-confidence 2D boxes can provide an additional strong supervision for the 3D student. However, in our previous pipeline, 2D teacher boxes can only supervise 3D indirectly through 3D teacher boxes that are potentially worse than 2D in terms of classification and localization. We propose to directly match 2D pseudo-labels and 3D student boxes and enforce box and class consistency between them as shown in Figure 6. Applying Hungarian Matching and thresholding as in Equation 7:

$$\left\{\left((\hat{\mathbf{b}}_{\mathbf{T}}, \hat{\mathbf{c}}_{\mathbf{T}})^{2D}, (\hat{\mathbf{b}}_{\mathbf{S}}, \hat{\mathbf{c}}_{\mathbf{S}})^{3D}\right)^{(j)}\right\}^{(<\tau_{hung})} = Hungarian^{\tau_{hung}}_{\text{2D-3D}}(\hat{\mathbf{y}}^{2D}_{\mathbf{T}}, \hat{\mathbf{y}}^{3D}_{\mathbf{S}}) \tag{11}$$

Then, the 2D-3D consistency loss between matched 2D and 3D pairs is:

$$\mathcal{L}_{consistency} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{iou}\mathcal{L}_{iou} + \lambda_{focal}\mathcal{L}_{focal} \tag{12}$$

Losses $\mathcal{L}_{L1}$ and $\mathcal{L}_{iou}$ are identical to the box consistency costs in Equation 8. $\mathcal{L}_{focal}$ is FocalLoss with 3D student probabilities supervised by the 2D teacher box class. This final 2D-3D consistency loss fully utilizes the strengths of RGB.

4 Experiments

4.1 Datasets and Evaluation Metrics

KITTI. We follow 3DIoUMatch and evaluate on the same 1% and 2% labeled frames sampled from 3712 training frames, and we also evaluate on 20% of driving *sequences*. We average over three splits for each % setting. We report for both 2D and 3D the moderate mAP for the Car, Pedestrian, and Cyclist classes.
Waymo Open Dataset. We also evaluate on the large-scale Waymo dataset, which has 158361 training frames. Each frame has 360 degree LiDAR and 5 RGB cameras, with the cameras only capturing 240 degrees. This limitation, coupled with the complex and diverse urban setting, makes multi-modal training especially difficult on Waymo. We validate our framework on the 1% labeled data setting, sampling 1% of the 798 sequences, which results in around 1.4k frames. Due to the sheer scale of the Waymo dataset and the observation that even this 1% split has four times the cars and eight times the pedestrians as the full KITTI dataset, we validate on a single Waymo split. We report mAP and mAPH at both LEVEL 1 and LEVEL 2 difficulties for Car and Pedestrian.

4.2 Implementation Details

We use PV-RCNN [48] for 3D detection and Faster-RCNN [45] with FPN [29] and ResNet50 [14] for 2D detection. To reduce labeling costs specifically for autonomous driving, we follow multi-modality methods [40, 52, 62] and pre-train the 2D detector on COCO [31]. This is a reasonable setting because labeling costs associated with annotating autonomous driving frames in 3D for specific applications do not preclude the existence of publicly available 2D detection datasets in another domain. Further, we find in Table 7 that DetMatch still dramatically improves over SSL baselines even without COCO pre-training.

We set $\tau_{3D} = 0.3$, $\tau_{2D} = 0.7$, and $\tau_{hung} = -1.5$, and use the same τ_{hung} threshold for both applications of Hungarian Matching. For KITTI, we train for 5k iterations with a batch size of 24; for Waymo, we train for 12k iterations with a batch size of 12. Additional details can be found in the supplementary.

4.3 Results on KITTI

We evaluate our model on 2D and 3D object detection on KITTI, comparing with 3DIoUMatch and our SSL baseline, which is equivalent to Unbiased Teacher [33]

Table 1: 3D detection performance comparison on KITTI. Training on the labeled samples to convergence, we observe slightly better labeled-only performance than 3DIouMatch. Improvement is increase from labeled-only results.

Method	1%				2%				20%			
	mAP	Car	Ped	Cyc	mAP	Car	Ped	Cyc	mAP	Car	Ped	Cyc
Labeled-Only (3DIoUMatch Reported)	43.5	73.5	28.7	28.4	54.3	76.6	40.8	45.5	-	-	-	-
3DIoUMatch	48.0	76.0	31.7	36.4	61.0	78.7	48.2	56.2	-	-	-	-
Improvement	+4.5	+2.5	+3.0	+8.0	+6.7	**+2.1**	+7.4	+10.7	-	-	-	-
Labeled-Only (Reproduced by Us)	45.9	73.8	30.4	33.4	55.8	76.1	44.9	46.4	61.3	77.9	47.1	58.9
Confidence Thresholding	54.4	75.9	42.7	44.6	63.3	76.5	50.0	63.4	68.1	77.8	58.0	68.6
Improvement	+8.5	+2.1	+12.3	**+11.2**	+7.5	+0.4	+5.1	+17.0	+6.8	-0.1	**+10.9**	+9.7
Ours	59.0	77.5	57.3	42.3	65.6	78.2	54.1	64.7	68.7	78.7	57.6	69.6
Improvement	**+13.1**	**+3.7**	**+26.9**	+8.9	**+9.8**	**+2.1**	**+9.2**	**+18.3**	**+7.4**	**+0.8**	+10.5	**+10.7**

Table 2: 2D detection performance comparison on KITTI. Note that although we train with projected 3D boxes, we evaluate with annotated 2D boxes.

Method	1%				2%				20%			
	mAP	Car	Ped	Cyc	mAP	Car	Ped	Cyc	mAP	Car	Ped	Cyc
Labeled-Only	65.3	86.6	68.6	40.8	68.9	87.4	70.7	48.3	63.9	87.5	64.5	39.8
Confidence Thresholding	60.4	86.1	69.2	25.8	65.5	87.6	71.5	37.2	66.2	88.8	70.0	39.7
Improvement	-4.9	-0.5	+0.6	-15.0	-3.4	+0.2	+0.8	-11.1	+2.3	+1.3	+5.5	-0.1
Soft-Teacher [71]	67.3	88.3	68.9	44.7	70.5	88.7	70.8	52.1	67.2	89.0	69.2	43.4
Improvement	+2.0	+1.7	+0.3	+3.9	+1.6	+1.3	+0.1	+3.8	+3.3	+1.5	+4.7	+3.6
Ours	71.4	88.8	73.9	51.7	74.5	89.0	74.6	59.9	72.8	89.1	71.6	57.7
Improvement	**+6.1**	**+2.2**	**+5.3**	**+10.9**	**+5.6**	**+1.6**	**+3.9**	**+11.6**	**+8.9**	**+1.6**	**+7.1**	**+17.9**

in 2D. The results are shown in Tables 1 and 2. First, we find that with a well-tuned 3D confidence threshold, our 3D-only confidence thresholding baseline is able to outperform 3DIoUMatch in both mAP absolute performance and improvement. However, we note that for the Car class, 3DIoUMatch outperforms the 3D SSL baseline which struggles to improve performance over labeled-only training in 2% and 20% settings. This is because Car is the most common class and is already well-trained just from the labeled data, making further improvements difficult. Our proposed DetMatch, leveraging both 2D and 3D detections, consistently outperforms all methods. Notably, we find that in the 1% setting, we observe a remarkable **26.9%** boost in AP, far outperforming 3DIoUMatch, which achieves a 3% improvement, and our 3D SSL baseline, which achieves a 12.3% improvement. This gap can be attributed to the ambiguity of pedestrians in 3D and the relative clarity of this class when viewed in the RGB image.

For 2D detection, we see that the Unbiased Teacher baseline suffers from a drop in performance through SSL training for 1% and 2% settings despite our hyperparameter search. Soft-Teacher [71] is able to improve performance, but only by a small margin. We attribute this to two factors. First, SSL on autonomous driving datasets is a more difficult setting than SSL on COCO because driving datasets like KITTI have less image diversity, making it more susceptible to over-fitting. Indeed, as the amount of labeled data increases for KITTI, 2D SSL

Table 3: Performance comparison on the validation set of the Waymo Dataset.

1% Data	3D								2D			
	Car L1		Car L2		Ped L1		Ped L2		Car		Ped	
	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	L1	L2	L1	L2
Labeled-Only	47.3	45.6	43.6	42.0	28.9	15.6	26.2	14.1	42.3	39.5	50.8	47.0
Confidence Thresholding	52.6	51.6	48.4	47.5	35.2	16.7	32.0	15.2	44.4	41.3	48.7	45.1
Improvement	**+5.3**	**+6.0**	**+4.8**	**+5.5**	+6.3	+1.1	+5.8	+1.1	+2.1	+1.8	-2.1	-1.9
Ours	52.2	51.1	48.1	47.2	39.5	18.9	35.8	17.1	47.8	44.4	50.6	46.8
Improvement	+4.9	+5.5	+4.5	+5.2	**+10.6**	**+3.3**	**+9.6**	**+3.0**	**+5.5**	**+4.9**	-0.2	-0.2

improves. We note that even the limited 1% setting on COCO has 1171 images, each in a completely different scene. On the other hand, KITTI 1% only has 37 images, and even the larger 20% setting, due to its constraint of sampling driving sequences, has comparatively lower scene diversity. These factors, coupled with pre-training on COCO which strengthens the original model, make improving on the labeled-only baseline difficult. Second, single-modality training is far more susceptible to self-training error propagation. Although the asymmetric augmentation and EMA work to decouple the student from the teacher, their predictions are still highly correlated, causing the student to overfit to its own predictions, including its own errors. Our results show that the proposed DetMatch is more robust to these factors, demonstrating substantial performance gains over the labeled-only and 2D SSL baselines. Notably, we find that detection of Cyclists, a rare category, declines by 15% mAP under Unbiased Teacher in KITTI 1% but improves by 10.9% mAP with DetMatch, a gap of 25.9% mAP.

4.4 Results on Waymo Open Dataset

To test the robustness of our framework, we additionally benchmark DetMatch on the difficult Waymo dataset. Because Waymo's 2D cameras have a combined FOV of 240 degrees, we use the 3D SSL pseudo-labels for the remaining 120 degrees when training DetMatch. We keep hyperparameters of DetMatch, which were tuned on KITTI, the same for Waymo and find that they are generally applicable. Our 3D and 2D results are summarized in Table 3. We find that the confidence thresholding baseline is strong, consistently demonstrating improvements of 5% or 6% on the mAP metric for 3D. For 2D, we see a smaller improvement and even observe the performance on pedestrian drop by two points. We attribute this to the same factors that caused a drop in KITTI - although Waymo dataset is larger, its 1% labeled data diversity less than that of COCO.

DetMatch slightly drops in performance for Cars in 3D compared to the SSL baseline. However, it improves on the SSL baseline by a substantial 4.3 mAP for Pedestrian L1. Further, DetMatch achieves a large boost of 3.4 mAP for Car L1 in 2D over single-modality SSL, and although it does not boost performance for 2D Pedestrian, DetMatch stymies the decline from Unbiased Teacher.

Overall, compared to the labeled-only and SSL baselines, our method significantly boosts performance for Pedestrian on 3D and Car on 2D while largely maintaining other settings' performance. We attribute the large Pedestrian 3D

Table 4: 3D Effect of τ_{hung}

3D Eval	mAP	Car	Ped	Cyc
Labeled-Only	45.9	73.8	30.4	33.4
$\tau_{hung} = -1$	54.2	76.1	49.3	37.2
$\tau_{hung} = -1.5$	**57.9**	76.7	**55.0**	**42.0**
$\tau_{hung} = -2$	52.4	**76.9**	43.7	36.7

Table 5: 2D Effect of τ_{hung}

2D Eval	mAP	Car	Ped	Cyc
Labeled-Only	65.3	86.6	68.6	40.8
$\tau_{hung} = -1$	69.3	87.9	70.4	49.5
$\tau_{hung} = -1.5$	**70.2**	88.7	**72.1**	**49.9**
$\tau_{hung} = -2$	56.5	**89.5**	52.3	27.7

Table 6: Ablation of DetMatch Modules

1% Data	3D				2D			
	mAP	Car	Ped	Cyc	mAP	Car	Ped	Cyc
Labeled-Only	45.9	73.8	30.4	33.4	65.3	86.6	68.6	40.8
+Confidence Thresholding	54.4	75.9	42.7	44.6	60.4	86.1	69.2	25.8
+ 2D-3D Teacher Matching	57.9	76.7	55.0	42.0	70.2	88.7	72.1	49.9
+ 2D Teacher & 3D Student Box Consistency	59.4	77.4	56.5	44.4	69.8	88.5	71.9	49.0
+ 2D Teacher & 3D Student Class Consistency	59.0	77.5	57.3	42.3	71.4	88.8	73.9	51.7
+ 2D Teacher & 3D Student MSE instead of Focal	58.2	77.6	57.7	39.3	68.1	88.6	72.0	50.8

improvement to DetMatch's effective use of RGB images' advantage in identifying and localizing this class. On the other hand, the Car 2D boost stems from the 2D detector benefiting from 3D's stronger detection of Cars, which are often highly occluded in the urban streets captured in Waymo. Thus, although our DetMatch does not uniformly boost all classes, perhaps due to Faster-RCNN with ResNet50 being an older and weaker model in 2D compared to PV-RCNN in 3D, the remarkable boost regardless in Pedestrian 3D detection and Car 2D detection demonstrate that our pipeline is effective in exploiting the unique advantages of each sensor to improve detections of the other modality.

4.5 Ablation Studies and Discussion

Here, we focus on quantitative results; visualizations are in the supplementary.

Threshold for DetMatch. Results for KITTI 1% at various τ_{hung} on DetMatch with just the 2D-3D Teacher Matching pseudo-labeling module are shown in Tables 4 and 5. Ablations on single-modality thresholds τ_{3D} and τ_{2D} are in the supplementary. We find that Car prefers a more stringent (lower) cost threshold. Further, we observe that 2D and 3D mAP both peak at the *same* $\tau_{hung} = -1.5$, which shows that improvements in one modality strongly benefit the other.

Ablation of Multi-Modal Components. Next, we study the effect of each module of DetMatch in Table 6. Components not part of our final model are in gray. We focus on the Car and Pedestrian classes for this fine-grained comparison as Cyclist results vary by up to 3 AP even on 100% labeled data runs. Replacing the single-modality thresholding with our 2D-3D teacher matched pseudo-labels results in a large improvement. This shows us that pseudo-labeling with objects consistently detected in both modalities better supervises the student.

Enforcing box consistency between the 2D teacher and 3D student improves substantially improves the 3D performance with a small 0.2 point drop in Car and Pedestrian 2D performance. We attribute this boost to the 3D student now generating boxes that better fit objects in the dense 2D image. FocalLoss class consistency boosts 3D and 2D Pedestrian performance by 0.8 and 2 points, respectively. This is in-line with our observations that Pedestrian is difficult to detect in 3D - by rectifying class prediction of under-confident or incorrect 3D detections using 2D, the 3D model improves. Further, the 2D performance improves because 2D pseudo-labels are tied with 3D teacher predictions. By train-

Table 7: Impact of COCO Pre-training

1% Data		3D				2D			
		mAP	Car	Ped	Cyc	mAP	Car	Ped	Cyc
w/ COCO Pre-Training	Labeled-Only	45.9	73.8	30.4	33.4	65.3	86.6	68.6	40.8
	Ours	59.0	77.5	57.3	42.3	71.4	88.8	73.9	51.7
w/o COCO Pre-Training	Labeled-Only	45.9	73.8	30.4	33.4	46.2	77.6	47.1	13.9
	Ours	57.1	77.7	55.3	38.3	59.1	85.9	59.0	30.7

ing the 3D model to generate more accurate 3D Pedestrian detections, the 2D model is better supervised as well. This improvement demonstrates the mutually beneficial relationship between improvements in the 2D and 3D models.

We try replacing FocalLoss in class consistency with MSE following Mean Teacher [60]. That this decreases performance gives us more insight into the purpose of class consistency. MSE encourages logit matching [21, 60], which is closely related to knowledge distillation [15], where, by imitating class similarities predicted by a teacher, the student learns the underlying function of the teacher. In our setting, the teacher and student are of different modalities and consume data of very different representations, inhibiting such mimicking. As such, what our consistency module does is directly interpretable - it rectifies 3D student box and class predictions using the 2D teacher outputs.

Without COCO Pre-training. We also evaluate our pipeline without COCO pre-training, as shown in Table 7. We find that although COCO pre-training is important for 2D performance, we still achieve strong 3D performance without it, notably maintaining a substantial 24.9% AP improvement for Pedestrian. This shows that DetMatch does not need COCO, instead benefiting more from the multi-modal interaction. Further, improvements from using COCO shows that our framework is a unique and effective way of transferring benefits from 2D labels, which are easier to annotate than 3D labels, to the 3D detection task.

5 Conclusion

In this work, we proposed DetMatch, a flexible multi-modal SSL framework for object detection that obtains state-of-the-art performance on various limited labeled data settings on KITTI and Waymo. We demonstrate that pseudo-labels generated by matching 2D and 3D detections allow each modality to benefit from the other's advantages and improvements. Further, by enforcing consistency between 3D student and 2D teacher boxes, we leverage the unique advantages that the dense RGB image gives the 2D detector in detecting ambiguous objects. As our pipeline achieves improved performance on 3D detection by using a COCO pre-trained 2D detector, our method also shows potential in leveraging cheaper or publicly available 2D annotations to lower 3D data requirements.

Acknowledgements: Co-authors from UC Berkeley were sponsored by Berkeley Deep Drive (BDD).

References

1. Bachman, P., Alsharif, O., Precup, D.: Learning with pseudo-ensembles. Advances in neural information processing systems **27** (2014)
2. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In: ICLR (2020)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems **32** (2019)
4. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. ArXiv **abs/2004.10934** (2020)
5. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 6154–6162 (2018)
6. Caine, B., Roelofs, R., Vasudevan, V., Ngiam, J., Chai, Y., Chen, Z., Shlens, J.: Pseudo-labeling for scalable 3d object detection. ArXiv **abs/2103.02093** (2021)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
8. Chong, Z., Ma, X., Zhang, H., Yue, Y., Li, H., Wang, Z., Ouyang, W.: Monodistill: Learning spatial features for monocular 3d object detection. ArXiv **abs/2201.10830** (2022)
9. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3070–3079 (2019)
10. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2432–2443 (2017)
11. Feng, D., Zhou, Y., Xu, C., Tomizuka, M., Zhan, W.: A simple and efficient multi-task network for 3d object detection and road understanding. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7067–7074. IEEE (2021)
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. 2012 IEEE Conference on Computer Vision and Pattern Recognition pp. 3354–3361 (2012)
13. Graham, B., Engelcke, M., Maaten, L.V.D.: 3d semantic segmentation with submanifold sparse convolutional networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 9224–9232 (2018)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)
15. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)
16. Huang, T., Liu, Z., Chen, X., Bai, X.: Epnet: Enhancing point features with image semantics for 3d object detection. In: ECCV (2020)
17. Janoch, A., Karayev, S., Jia, Y., Barron, J.T., Fritz, M., Saenko, K., Darrell, T.: A category-level 3-d object dataset: Putting the kinect to work. In: ICCV Workshops (2011)

18. Jaritz, M., Vu, T.H., de Charette, R., Wirbel, É., Pérez, P.: xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12602–12611 (2020)
19. Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. In: NeurIPS (2019)
20. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 784–799 (2018)
21. Kim, T., Oh, J., Kim, N., Cho, S., Yun, S.Y.: Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. In: IJCAI (2021)
22. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**, 83–97 (1955)
23. Lahoud, J., Ghanem, B.: 2d-driven 3d object detection in rgb-d images. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 4632–4640 (2017)
24. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR (2017)
25. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013)
26. Li, H., Wu, Z., Shrivastava, A., Davis, L.S.: Rethinking pseudo labels for semi-supervised object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1314–1322 (2022)
27. Li, Y.J., Park, J., O'Toole, M., Kitani, K.: Modality-agnostic learning for radar-lidar fusion in vehicle detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 918–927 (2022)
28. Liang, Z., Zhang, M., Zhang, Z., Zhao, X., Pu, S.: Rangercnn: Towards fast and accurate 3d object detection with range image representation. ArXiv **abs/2009.00206** (2020)
29. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 936–944 (2017)
30. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 318–327 (2020)
31. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
32. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV (2016)
33. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. ICLR (2021)
34. Liu, Y.C., Huang, Y.K., Chiang, H., Su, H.T., Liu, Z.Y., Chen, C.T., Tseng, C.Y., Hsu, W.H.: Learning from 2d: Pixel-to-point knowledge transfer for 3d pretraining. ArXiv **abs/2104.04687** (2021)
35. Liu, Y., Yi, L., Zhang, S., Fan, Q., Funkhouser, T.A., Dong, H.: P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding. ArXiv **abs/2012.13089** (2020)
36. Liu, Z., Qi, X., Fu, C.W.: 3d-to-2d distillation for indoor scene parsing. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4462–4472 (2021)

37. Park, J.D., Weng, X., Man, Y., Kitani, K.: Multi-modality task cascade for 3d object detection. BMVC (2021)
38. Qi, C., Chen, X., Litany, O., Guibas, L.: Imvotenet: Boosting 3d object detection in point clouds with image votes. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4403–4412 (2020)
39. Qi, C., Litany, O., He, K., Guibas, L.: Deep hough voting for 3d object detection in point clouds. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9276–9285 (2019)
40. Qi, C., Liu, W., Wu, C., Su, H., Guibas, L.: Frustum pointnets for 3d object detection from rgb-d data. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 918–927 (2018)
41. Qi, C., Yi, L., Su, H., Guibas, L.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: NIPS (2017)
42. Qi, C., Zhou, Y., Najibi, M., Sun, P., Vo, K.T., Deng, B., Anguelov, D.: Offboard 3d object detection from point cloud sequences. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6130–6140 (2021)
43. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. Advances in neural information processing systems **28** (2015)
44. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 779–788 (2016)
45. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 1137–1149 (2015)
46. Rezatofighi, S.H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I.D., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 658–666 (2019)
47. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Advances in neural information processing systems **29** (2016)
48. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10526–10535 (2020)
49. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–779 (2019)
50. Shi, S., Wang, Z., Shi, J., Wang, X., Li, H.: From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. IEEE transactions on pattern analysis and machine intelligence (2020)
51. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
52. Sindagi, V., Zhou, Y., Tuzel, O.: Mvx-net: Multimodal voxelnet for 3d object detection. 2019 International Conference on Robotics and Automation (ICRA) pp. 7276–7282 (2019)
53. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems **33**, 596–608 (2020)

54. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. ArXiv **abs/2005.04757** (2020)
55. Song, G., Liu, Y., Wang, X.: Revisiting the sibling head in object detector. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11560–11569 (2020)
56. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 567–576 (2015)
57. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S.M., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2443–2451 (2020)
58. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., Luo, P.: Sparse r-cnn: End-to-end object detection with learnable proposals. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14449–14458 (2021)
59. Tang, Y., Chen, W., Luo, Y., Zhang, Y.: Humble teachers teach better students for semi-supervised object detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3131–3140 (2021)
60. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
61. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9626–9635 (2019)
62. Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4603–4611 (2020)
63. Wang, C.H., Chen, H.W., Fu, L.C.: Vpfnet: Voxel-pixel fusion network for multi-class 3d object detection. ArXiv **abs/2111.00966** (2021)
64. Wang, H., Cong, Y., Litany, O., Gao, Y., Guibas, L.J.: 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14610–14619 (2021)
65. Wang, J., Gang, H., Ancha, S., Chen, Y.T., Held, D.: Semi-supervised 3d object detection via temporal graph neural networks. 2021 International Conference on 3D Vision (3DV) pp. 413–422 (2021)
66. Wang, Z., Jia, K.: Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp. 1742–1749 (2019)
67. Xiao, J., Owens, A., Torralba, A.: Sun3d: A database of big spaces reconstructed using sfm and object labels. 2013 IEEE International Conference on Computer Vision pp. 1625–1632 (2013)
68. Xie, L., Xiang, C., Yu, Z., Xu, G., Yang, Z., Cai, D., He, X.: Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module. AAAI **abs/1911.06084** (2020)
69. Xu, C., Yang, S., Zhai, B., Wu, B., Yue, X., Zhan, W., Vajda, P., Keutzer, K., Tomizuka, M.: Image2point: 3d point-cloud understanding with pretrained 2d convnets. arXiv preprint arXiv:2106.04180 (2021)

70. Xu, C., Zhai, B., Wu, B., Li, T., Zhan, W., Vajda, P., Keutzer, K., Tomizuka, M.: You only group once: Efficient point-cloud processing with token representation and relation inference module. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4589–4596. IEEE (2021)
71. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3040–3049 (2021)
72. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors (Basel, Switzerland) **18** (2018)
73. Yang, Q., Wei, X., Wang, B., Hua, X., Zhang, L.: Interactive self-training with mean teachers for semi-supervised object detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5937–5946 (2021)
74. Yang, Z., Sun, Y., Liu, S., Jia, J.: 3dssd: Point-based 3d single stage object detector. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11037–11045 (2020)
75. Yin, T., Zhou, X., Krähenbühl, P.: Multimodal virtual point 3d detection. NeurIPS (2021)
76. Yoo, J.H., Kim, Y., Kim, J.S., Choi, J.W.: 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In: ECCV (2020)
77. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Advances in Neural Information Processing Systems **34**, 18408–18419 (2021)
78. Zhang, H., Cissé, M., Dauphin, Y., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. ICLR (2018)
79. Zhao, L., Zhou, H., Zhu, X., Song, X., Li, H., Tao, W.: Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation. ArXiv **abs/2108.07511** (2021)
80. Zhao, N., Chua, T.S., Lee, G.H.: Sess: Self-ensembling semi-supervised 3d object detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11076–11084 (2020)
81. feng Zhou, Q., Yu, C., Wang, Z., Qian, Q., Li, H.: Instant-teaching: An end-to-end semi-supervised object detection framework. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4079–4088 (2021)
82. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 4490–4499 (2018)